DIGITAL PRESERVATION FOR INSTITUTIONAL REPOSITORIES

ELIZABETH LA BEAUD

DIGITAL LAB MANAGER, UNIVERSITY OF SOUTHERN MISSISSIPPI

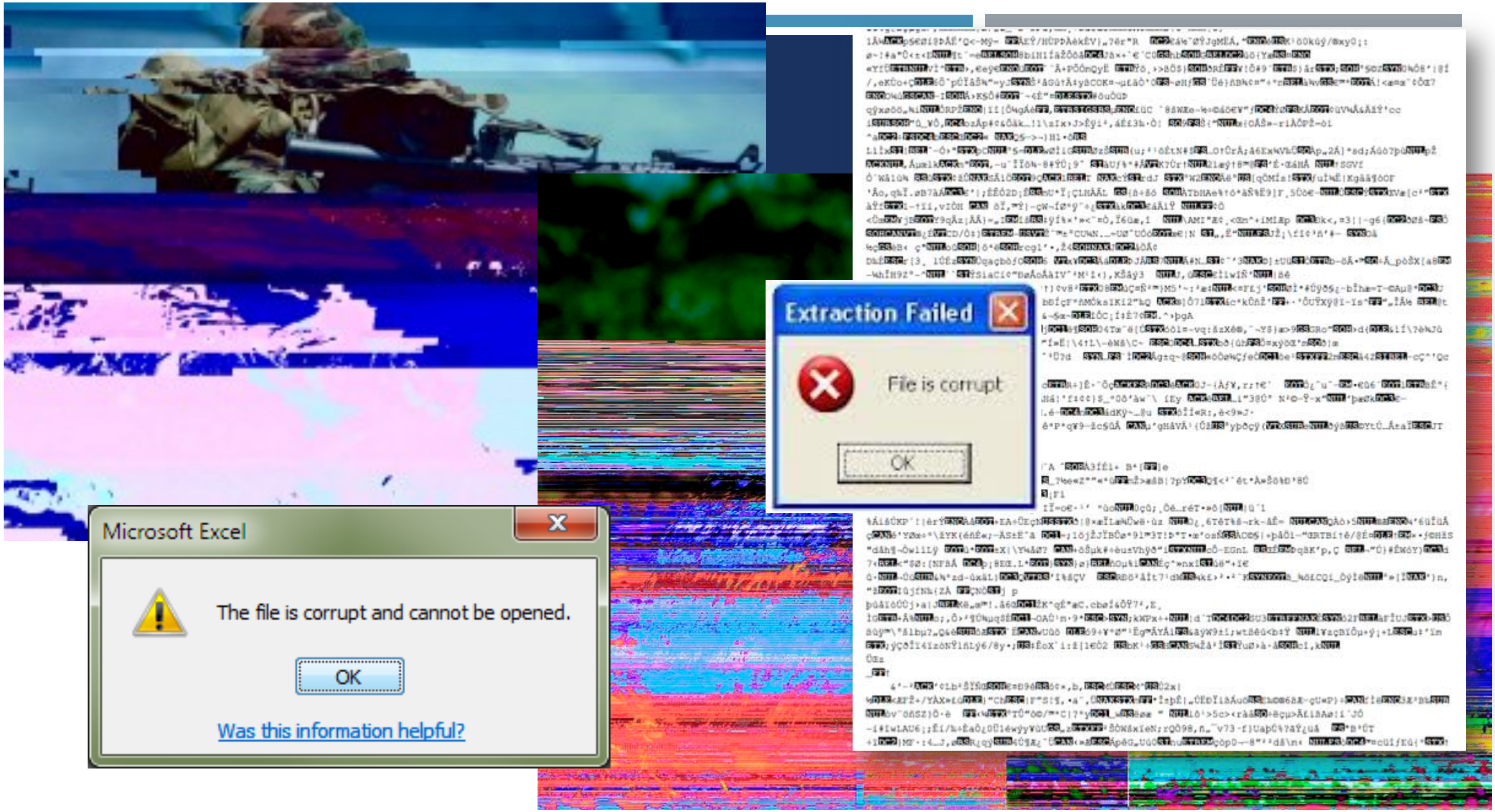
ACCESS VS. PRESERVATION

ACCESS

- relies on **cutting edge** technologies to provide best and fastest access at a point in time
- **selects** metadata needed to use and understand content
- access systems **deliver** objects with user-oriented services to make the objects
- purpose: **provide content to users**
- focus**: current users**

PRESERVATION

- relies upon **proven** technologies to preserve digital objects across generations of technology
- **accumulates** metadata over the life cycle to trace and preserve content
- preservation systems **create** new versions of digital objects for access to deliver as needs change over time
- purpose: **ensure long-term access**
- focus: **future** users


Extraction Failed
File is corrupt
OK
Microsoft Excel
The file is corrupt and cannot be opened.
OK
Was this information helpful?

RISKS TO DIGITAL CONTENT

- Obsolescence – hardware & software
- Loss of power
- Hardware failure
- Change and loss – accidental and intentional
- Bit rot
- User error
- Man-made and natural disasters
- Cyber attacks
- Format obsolescence
- Media failure
- Inappropriate access (i.e. confidential data)
- Non-compliance – standards and requirements
- Dust

PRESERVATION MAKES LONG-TERM ACCESS POSSIBLE…

NDSA'S LEVELS OF DIGITAL PRESERVATION

Table 1: Version 1 of the Levels of Digital Preservation

	Level 1 (Protect your data)	Level 2 (Know your data)	Level 3 (Monitor your data)	Level 4 (Repair your data)
Storage and Geographic Location	- Two complete copies that are not collocated - For data on heterogeneous media (optical discs, hard drives, etc.) get the content off the medium and into your storage system	- At least three complete copies - At least one copy in a different geographic location - Document your storage system(s) and storage media and what you need to use them	- At least one copy in a geographic location with a different disaster threat - Obsolescence monitoring process for your storage system(s) and media	- At least three copies in geographic locations with different disaster threats - Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems
File Fixity and Data Integrity	- Check file fixity on ingest if it has been provided with the content - Create fixity info if it wasn't provided with the content	- Check fixity on all ingests - Use write-blockers when working with original media - Virus-check high risk content	- Check fixity of content at fixed intervals - Maintain logs of fixity info; supply audit on demand - Ability to detect corrupt data - Virus-check all content	- Check fixity of all content in response to specific events or activities - Ability to replace/repair corrupted data - Ensure no one person has write access to all copies
Information Security	- Identify who has read, write, move and delete authorization to individual files - Restrict who has those authorizations to individual files	- Document access restrictions for content	- Maintain logs of who performed what actions on files, including deletions and preservation actions	- Perform audit of logs
Metadata	- Inventory of content and its storage location - Ensure backup and non-collocation of inventory	- Store administrative metadata - Store transformative metadata and log events	- Store standard technical and descriptive metadata	- Store standard preservation metadata
File Formats	- When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs	- Inventory of file formats in use	- Monitor file format obsolescence issues	- Perform format migrations, emulation and similar activities as needed

You have options:

- Buy, build, or join
- Local installation

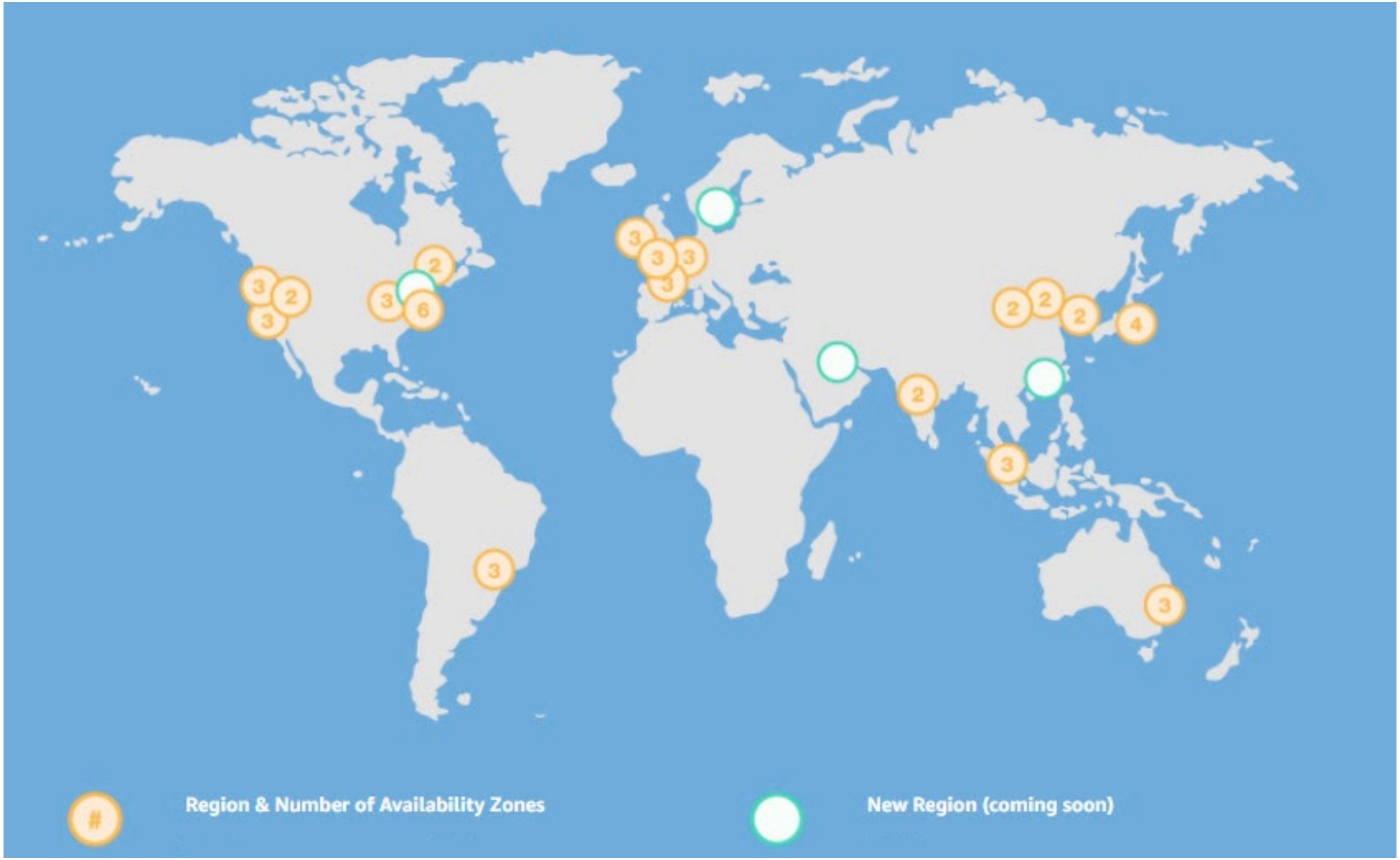
- Open source
- Hosted solutions
- Proprietary

CONSIDER…

- Cost (available resources for preservation)
- Quantity (size and number of files)
- Expertise (skills required to manage)
- Partnerships (geographic distribution)
- Services (outsourcing)

WHERE IS YOUR DATA?

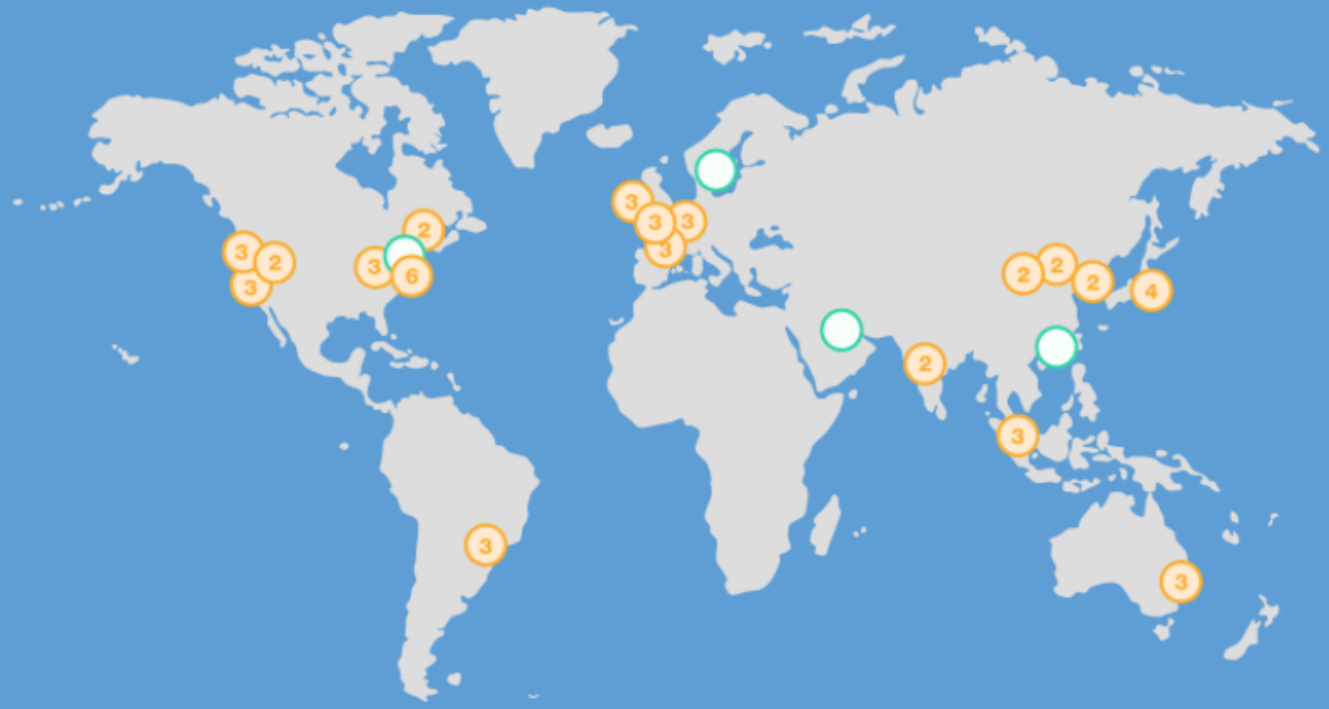
WHERE IS THE CLOUD?


Region & Number of Availability Zones
New Region (coming soon)

Amazon Web Services

Global Network of Regions and Edge Locations

The AWS Cloud spans 52 Availability Zones within 18 geographic Regions around the world, with announced plans for 12 more Availability Zones and four more Regions in Bahrain, Hong Kong SAR, Sweden, and a second AWS GovCloud Region in the US.



Region & Number of Availability Zones

US East
N. Virginia (6), Ohio (3)

US West
N. California (3), Oregon (3)

Asia Pacific
Mumbai (2), Seoul (2), Singapore (3), Sydney (3), Tokyo (4)

Canada
Central (2)

China
Beijing (2), Ningxia (2)

Europe
Frankfurt (3), Ireland (3), London (3), Paris (3)

South America
São Paulo (3)

AWS GovCloud (US-West) (2)

New Region (coming soon)

Bahrain

Hong Kong SAR, China

Sweden

AWS GovCloud (US-East)

AT LEAST TWO COPIES IN TWO LOCATIONS…

Storage and Geographic Location	- Two complete copies that are not collocated - For data on heterogeneous media (optical discs, hard drives, etc.) get the content off the medium and into your storage system	- At least three complete copies - At least one copy in a different geographic location - Document your storage system(s) and storage media and what you need to use them	- At least one copy in a geographic location with a different disaster threat - Obsolescence monitoring process for your storage system(s) and media	- At least three copies in geographic locations with different disaster threats - Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems

PLAN FOR STORAGE SYSTEM REPLACEMENT

WHO HAS ACCESS TO YOUR DATA?

WHAT HAPPENS WHEN THEY LEAVE?

SECURITY

Information Security	- Identify who has read, write, move and delete authorization to individual files - Restrict who has those authorizations to individual files	- Document access restrictions for content	- Maintain logs of who performed what actions on files, including deletions and preservation actions	- Perform audit of logs

Make sure

no *one person*

has write privileges

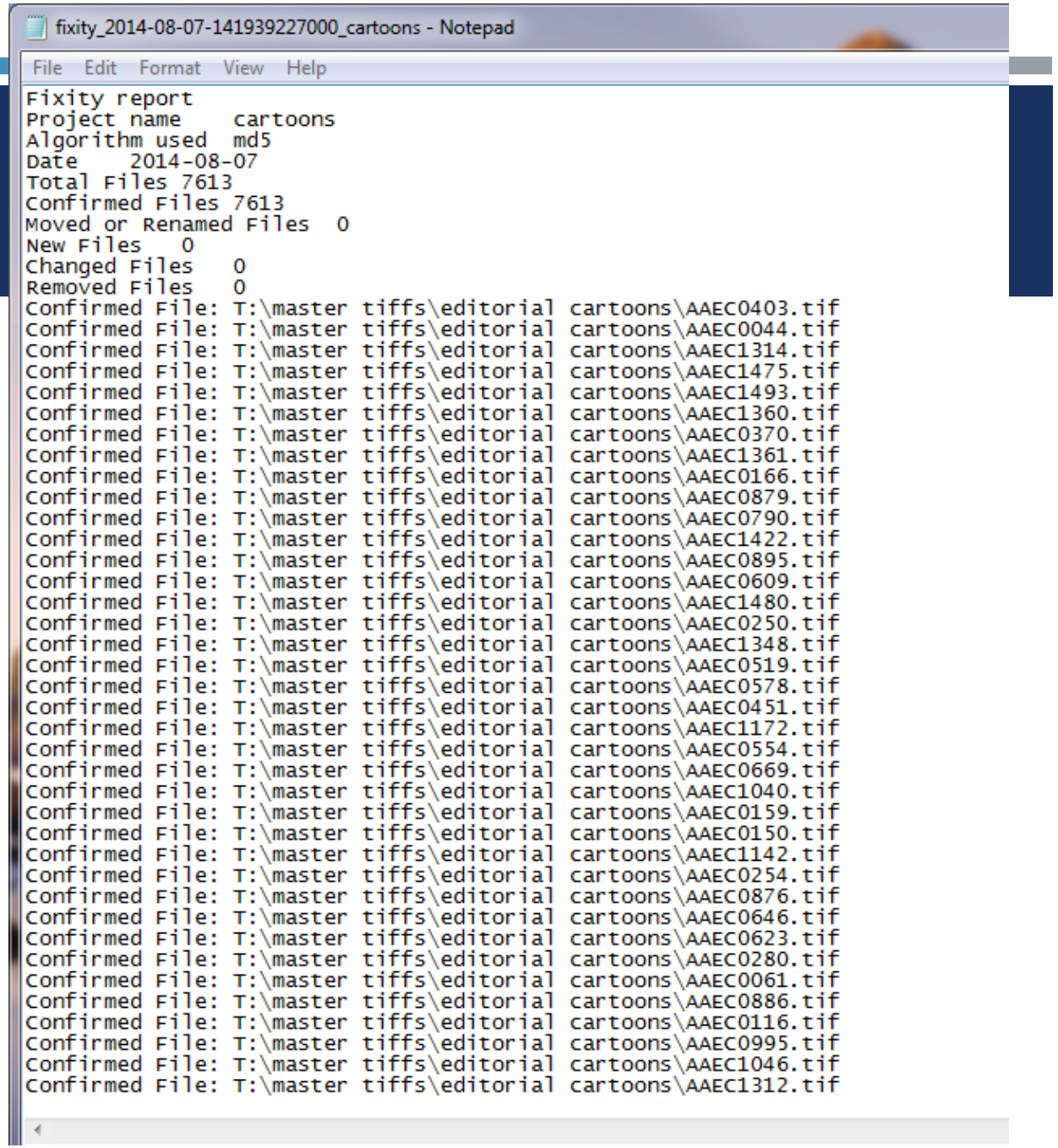
to all copies

FILE FIXITY & DATA INTEGRITY

File Fixity and Data Integrity	- Check file fixity on ingest if it has been provided with the content - Create fixity info if it wasn't provided with the content	- Check fixity on all ingests - Use write-blockers when working with original media - Virus-check high risk content	- Check fixity of content at fixed intervals - Maintain logs of fixity info; supply audit on demand - Ability to detect corrupt data - Virus-check all content	- Check fixity of all content in response to specific events or activities - Ability to replace/repair corrupted data - Ensure no one person has write access to all copies

ALERTS YOU TO A PROBLEM

- Checksums
  - SHA256
  - MD5
- MD5 Summer
- AVPreserve's Fixity
- File Verifier ++
  - (these are windows examples)



CAN YOUR VENDOR PROVE INTEGRITY ON DEMAND?

HOW OFTEN DO THEY CHECK FILE INTEGRITY?

FILE FORMAT SUSTAINABILITY

File Formats	- When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs	- Inventory of file formats in use	- Monitor file format obsolescence issues	- Perform format migrations, emulation and similar activities as needed

WHAT ARE YOU ACCEPTING?

HOW ARE YOU GOING TO MIGRATE IT OVER TIME?

TIERED OPTIONS

- **Tier 1**
- Open Sustainable File Formats (e.g., .txt)
- We will preserve your data to the best of our ability, and migrate the content over time to ensure accessibility. We'll keep a copy of your original file as well, just in case you want to revisit the original bitstream (1's and 0's).
- **Tier 2**
- Some well-known, but proprietary files (e.g., Microsoft Word)
- We will preserve your original file, but are limited in how we can maintain it. We'll always keep the 1's and 0's in the right order, but some formatting or other features may change.
- **Tier 3**
- Proprietary file formats with little to no open documentation (e.g., .spss)
- There are times when a specialized format is required for your data. In these cases, we will work with you. We will commit to preserving the original bitstream (1's and 0's) and the submitted metadata; however, while the files may or may not be renderable in the future, we can make no guarantees on the form or presentation it will take.

SUSTAINABLE FILE FORMAT CHARACTERISTICS

- Widely used by appropriate research community
- Non-proprietary
- Well documented, open standard
- Uncompressed (size permitting)
- Unencrypted
- Encoded in standard representation (e.g., ASCII, UTF-8)

METADATA

Metadata	- Inventory of content and its storage location - Ensure backup and non-collocation of inventory	- Store administrative metadata - Store transformative metadata and log events	- Store standard technical and descriptive metadata	- Store standard preservation metadata

DESCRIBE IT SO YOU CAN FIND IT!

METADATA FIELDS

- Identifier
- Title
- Description
- Creator
- Department
- Legacy department
- Date
- coverage
- Subjects
- Geographic location
- Language
- Type
- File format
- Contributors
- Permissions
- Disciplines
- Related docs

RECORDS CHANGES OVER TIME

- Document metadata changes
- File format migrations
- IP problems & resolutions
- Deaccessions

DOCUMENT, DOCUMENT, DOCUMENT…

POLICIES & PROCEDURES

TRUCK BOOK…

IT'S A COMMITMENT…

IT'S NOT A ONE AND DONE…

BUT…YOU DON'T HAVE TO DO IT ALONE…

RESOURCES

THANK YOU !!

ELIZABETH LA BEAUD

ELIZABETH.LABEAUD@USM.EDU